91602
new 82-10

ontario hydro
research division

COMPLIMENTARY COPY
Please Credit Ontario Hydro Research if
Reproduced in Whole or in Part

To: Dr. D. Mills
Director of Research

OUTLIERS IN THE NATIONAL METER SAMPLING PLAN

F. Camacho

A policy is proposed for dealing with outlier observations in samples obtained through the National Sampling Plan for revenue meters. This policy is needed to enhance and validate the applicability of the samples. The policy defines outlier observations, describes how to detect them, and determines their effect on accepting or rejecting a population.

1. INTRODUCTION

To enhance the applicability of the National Sampling Plan/1/, it is necessary to outline a policy for dealing with outliers in the samples. The goals of this policy are the following.

1. Define outlier observations and give a procedure to detect them.

2. Describe what to do with outliers in the analysis of the sample results. In particular, determine whether or not a population should be rejected based on the proportion of outliers.

The next Section discusses some of the theoretical issues associated with the problem of outliers; based on this discussion, Section 3 presents the outlier policy. Finally, Section 4 gives some concluding remarks.

2. THEORETICAL CONSIDERATIONS

In general, an observation is considered *outlier* if it is "sufficiently unlikely [to occur] under the assumed probability model"/2/. When the probability distribution of the data is completely specified, unlikely observations can be defined as values having a very small probability of occurrence, for example 10^{-4} or 10^{-5}. However, in practical applications finite samples instead of theoretical distributions are available and in this case it is not clear how to define "sufficiently unlikely observations". Reference/2/ suggests the following two step procedure.

job	file	date	report no.
72-510-000	809.7	August 19, 1987	87-205-K

1. Identify the *outlying values*. These are defined as those exhibiting a different behaviour compared to that of the majority of the data.

2. Examine each outlying values to decide if it is an outlier.

2.1 Identification of Outlying Values

The outlying values are the observations outside a given interval $[a, b]$, where a and b are selected to reflect the location of the majority of the data. One could suggest the interval $[x - 3s, x + 3s]$, where x is the sample mean and s is the sample standard deviation, as the interval containing the majority of the data. However, this is not practical because the limits of the interval are significantly affected by the outliers. The limits a and b must be define so that they are not influenced by the magnitude of the outliers. This can be achieved in the following way (see /2/).

(i) Arrange the n available observations $x_1, \ldots, x_n$ in ascending order. Denote the result as $x_{(1)}, \cdots, x_{(n)}$.

(ii) Calculate the *depth* of each observation $x_{(i)}$, $i = 1,\ldots,n$, as the minimum between i and $n + 1 - i$. If the bracket $[x]$ denote the operation of finding the largest integer less or equal to x, then the maximum depth value is $[(n+1)/2]$. Note that there are two observations associated with a depth value $d = 1,2,\ldots,[n/2]$: a lower value, $x_{(d)}$, and an upper value, $x_{(n+1-d)}$. It is also possible to associate two observations to depth values of the form $d + 0.5$: a lower value equal to $(x_{(d)} + x_{(d+1)})/2$ and a upper value equal to $(x_{(n-d)} + x_{(n+1-d)})/2$.

(iii) Calculate the *median* of the sample as the observation with depth $(n + 1)/2$. If n is even, $n = 2k$, the median falls halfway between $x_{(k)}$ and $x_{(k+1)}$; if n is odd, $n = 2k + 1$, the median is $x_{(k)}$. Summarizing,

$$\text{median} = \begin{cases} (x_{(k)} + x_{(k+1)})/2 & \text{for } n = 2k. \\ x_{(k)} & \text{for } n = 2k + 1. \end{cases}$$

(iv) Calculate the *fourths* of the sample as the two points with depth defined by

depth of fourth = ([depth of median] + 1)/2.

Denote the lower fourth as F_L and the upper fourth as F_U.

(v) Calculate the *fourth-spread*, d_F, as

$$d_F = F_U - F_L.$$

(vi) Calculate the limits of the interval $[a, b]$ as $F_L - 1.5d_F$ and $F_U + 1.5d_F$.

2.2 Determination of Outliers

The decision to declare an outlying value as an outlier should depend on a careful examination of the value to detect whether or not unusual factors affected the observation. To illustrate the fact that outlying values should not be declared outliers automatically, consider the simulation experiments carried out in reference/3/. By applying the procedure in Sec-

tion 2.1, they found that samples of size n, *without outliers*, taken from a population with a theoretical normal distribution had on average $0.4 + 0.007n$ outlying values. This implies that samples of 500 items will have about four outlying values, although none of these need to be an outlier.

A policy for detecting outliers in the meter samples is given in the next Section.

3. OUTLIER POLICY

3.1 Detection of Outliers

Observations from the sample should be declared outliers according to the following procedure.

1. All the meters in the sample should be included in this first screening process. Broken and/or stopped meters should be recorded with 100% Total Registration Error, TRE.

2. Using the *complete set* of meters, calculate F_L, F_U, and d_F as described in Section 2.

3. Declare all meters outside the interval $[F_L - 2.5d_F , F_U + 2.5d_F]$ as outliers.

4. Of the remaining meters, declare those outside the interval $[F_L - 1.5d_F , F_U + 1.5d_F]$ as outlying values.

5. Inspect all the outlying values and declare those that were subject to unusual factors (for example broken or stopped) as outliers.

3.2 The Effect of Outliers on Accepting/Rejecting the Population

The decision of what to do with the outliers depends on the proportion of outliers. If the sample does not contain a large proportion of outliers, they should be removed from further analyses. On the other hand, if the sample contains a large proportion of outliers and those are outside the interval [-3%, 3%] TRE, the population should be rejected. In fact, there would be enough evidence to reject the lot based on a sample by attributes.

The removal of outliers from the sample can be seen as a way of introducing bias in the analysis because it removes only the largest observations. It has been the experience of the utilities that these observations arise from meters that are broken prior of removal or that are damaged at the time of removal. However, these observations can be avoided. In fact, the former can be detected during routine readings and therefore, they can be replaced at any moment; the latter can be avoided if more care is taken in the handling. Therefore, the outliers should not be nuisance values for the analysis. However, if the proportion of stopped meters in the sample is too high, there is an indication that this is also the case in the population, implying that the population may not be conforming with the DCCAC requirements.

To decide whether or not to reject the population based on the proportion of outliers in the sample, the following procedure can be used. It is assumed that the sample specifies a consumers risk of 1% for population with more than 1% proportion of meters outside the interval [-3%,3%] TRE.

1. Calculate n_o, the number of outliers in a sample of size n, and z_o as

$$z_o = (n_o - 0.01n - 0.5)/\sqrt{(0.01)(0.99)n}\ .$$

2. Calculate p_o, defined as the probability that a standard normal variable be larger or equal to z_o. If $p_o < 0.01$ then reject the population. Otherwise, remove the outliers and decide the acceptance or rejection of the population using the triangular region described in/1/.

4. DISCUSSION

The new sampling policies outlined by the Department of Consumers and Corporate Affairs, DCCA/4/ consider high load, HL, and low load, LL, readings independently. However, the detection of outliers should consider both readings at the same time. The way to do this is as follows: (i) detect HL outliers using the HL reading; (ii) detect LL outliers using the LL reading; (iii) denote as ouliers the meters that are either HL or LL outliers. These meters should then be used to decide the acceptance/rejection of the population based on outliers. If the population is accepted, they should be removed from the sample before continuing with the analysis.

A new graphical procedure has recently been proposed in the statistical literature/5/. The method seems to be more automatic and accurate than the described in the report. However, the new method is quite intricate and requires advanced concepts for its understanding. It is belived that the method provided in this report is more understandable and its performance is quite satisfactory.

Approved:

J.G. Cassan
Manager
Operations Research Department

Submitted:

F. Camacho
Mathematician
Reliability & Statistics Section

FC:cmh

REFERENCES

1. Camacho, F., and Funhani, A.L., (1987). "Development of a National Sampling Plan for Revenue Meters for Canadian Utilities." Canadian Electrical Association Report No. 507 U 526.

2. Hoagling, D.C., (1983). "Letter Values: A Set of Selected Order Statistics." In *Understanding Robust & Exploratory Data Analysis*, Edited by Hoagling, D.C., Mosteller, F. & Tukey, J.W. John Wiley & Sons, Inc., New York.

3. Hoagling, D.C., Iglewicz, B., and Tukey. J.W., (1981). "Small-Sample Performance of a Resistant Rule for Outlier Detection." *1980 Proccedings of the Statistical Computing Section, Washington, D.C., American Statistical Association*, pp. 148-152.

4. Consumers and Corporate Affairs Canada, (1986). "Statistical Sampling Plans for the Verification and Reverification of Electricity and Gas Meters." Legal Metrology Branch Report No. LMB-EG-04.

5. Bacon-Shone, J. and Fung, W.K., (1987). "A new Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data." *Applied Statistics*, Vol. 36 pp. 153-162.

DISTRIBUTION

Dr. D. Mills	Director of Research
--Mr. F. Mason	Consumer & Corporate Affairs Canada Legal Metrology Branch, Holland Avenue Ottawa, Ontario, K1A 0C9
--Mr. R. Foisy	Service Electrometrie Direction Distribution Hydro-Quebec 201 rue Jarry ouest Montreal, Quebec H2P 1S7
--Mr. H. Douglas	Canadian Electrical Association Suite 580, 1 Westmount Square Montreal, Quebec H3Z 2P9
--Mr. W.M. Armstrong	Nova Scotia Power Corporation PO Box 910, Halifax, Nova Scotia
Mr. G. Sanchioni	Standard Products Department, KD200
Mr. S.J. Peukurinen	Standard Products Department, KD200
Mr. J. Cassan	Operations Research, KR297
Dr. A.L. Punhani	Operations Research, KR297
Dr. F. Camacho	Operations Research, KR297

Public Reference Centre
Editorial Section
Office Copies (2)
Extra Copies (2)